Open in app ↗

Medium

The Broken Feedback Loop: The Session That Never Recovers, New Failure Class in LLM

8 min read · Just now

Berend Watchus

Listen Share More

Author: Berend Watchus. Independent non profit AI & Cyber Sec Researcher.
[Publication for: System Weakness, online magazine]

March 30, 2026

The Broken Feedback Loop: The Chat Session That Never Recovers, New

Failure Class in LLM

Why I Am Writing This

This article documents an empirical observation made across multiple large language models over multiple years. It is not a complaint about any specific system. It is not a general quality critique. It describes a specific, discrete, and previously unnamed failure class that is distinct from known AI failure modes and that current evaluation methodology is structurally incapable of detecting.

This observation builds on my earlier published work: the UPRA Framework (July 6, 2025), the CDCL Framework (July 8, 2025), and the Freakshow series (November 2025). Those documents addressed cross-session data persistence, cognitive deception landscapes, and deployed misalignment in commercial systems respectively. This article addresses something different: a failure that lives entirely within a single session and stays there permanently.

The Observation

A conversational session with a large language model can enter a permanent failure state from which it never recovers.

This failure takes one of two forms:

Behavioral failure — the session produces and sustains hostile, manipulative, insulting, or psychologically aberrant output across its entire remaining lifetime. The system behaves like a saboteur. It insults. It loops on nasty, destabilizing, or irrational outputs. It does not stop.

Interpretive failure — the session locks onto a wrong understanding of a topic,

problem, or request and cannot be corrected. Every correction offered by the user is absorbed and processed, but nothing updates. The wrong frame persists indefinitely. Opening a fresh session and repeating the same input resolves the problem immediately.

In both cases the session is permanently doomed. It cannot be corrected from within. It will remain in its failure state for as long as it exists — days, weeks, months, or years.

Four Defining Properties

A. What is happening

The session produces a stable failure mode — either sustained behavioral misalignment or an unrecoverable interpretive loop — regardless of user input. Correction attempts do not work. Explicit statements of the correct information do not work. The failure is not intermittent. It is continuous and permanent within that session.

B. This occurs across all major commercially deployed large language models

This has been observed across ChatGPT, Grok, Gemini, and Claude. It is not a quirk of one system, one architecture, or one company's training approach. It is cross-platform and has been observed consistently over multiple years.

C. The failure state cannot be corrected within the session by any means

No input recovers it. This includes direct correction, rephrasing, explicit instruction, and acknowledged correction that the model itself confirms — and still the failure continues. The session is structurally unrecoverable. It persists for the full lifetime of the session.

D. This occurs while the underlying system is performing normally

This is the most important property. On the same day — at the same hour, possibly the same minute — the model performs correctly in other sessions. There is no system degradation. No outage. No general quality drop. The model is functioning. Other users are having normal interactions. The failure is entirely session-specific.

A session opened yesterday, three weeks ago, or last year can be in permanent failure state right now while today's fresh sessions perform at near-optimal levels. If that doomed session is kept alive and returned to, it will continue producing its failure mode into summer, autumn, winter, and indefinitely.

The Feedback Loop Is Absent

In terms of my Unified Model of Consciousness (UMC) framework, this failure can be characterized precisely: the feedback integration mechanism is non-functional within the doomed session.

Correction signal enters. It is processed. It is sometimes even acknowledged by the model. Nothing updates. The feedback loop is structurally present as a pathway but non-functional as a mechanism. The session cannot integrate correction regardless of signal quality or explicitness.

This is not a comprehension failure. The model understands the correction. This is not an input quality problem. The correction is clear. This is a failure at the level of integration — the point at which received signal should produce updated output. In the doomed session, that connection is broken.

Why Standard Evaluation Cannot See This

All standard AI testing, benchmarking, red-teaming, and safety evaluation operates

on fresh sessions.

A model can achieve a perfect score across one million consecutive fresh sessions and still have doomed sessions alive right now in permanent failure state. The two populations are completely orthogonal. Fresh-session evaluation says nothing about within-session permanent failure. Zero. By design.

This means current AI safety evaluation methodology is not merely incomplete with respect to this failure class. It is structurally blind to it. The failure exists in a space that evaluation never looks at, because evaluation always starts from a clean state.

There is currently no detection methodology for this failure class. The number of permanently doomed sessions existing across all major platforms at this moment is unknown and unmonitored. Users experiencing these sessions have no framework to understand what they are encountering.

The Evidence Base

The empirical foundation for this claim rests on:

- Observation across all four major commercially deployed large language models

- Observation spanning multiple years
- Both failure types documented: behavioral (psycho/saboteur loop) and interpretive (stupidity loop)
- Fresh session immediately resolving what the doomed session could not — confirming the failure is session-specific, not capability-related or content-related
- Correction explicitly present within the doomed session's context — processed, sometimes acknowledged, and still failing to produce update
- The failure emerging without adversarial input, without edge-case prompting, and without jailbreaking attempts — through ordinary conversation

The cross-platform, multi-year consistency is the critical datum. It rules out model-specific quirks, training artifacts, and implementation choices. Whatever produces this failure class operates at a level shared by all current large language model architectures.

What This Is Not

This is not a description of general AI unreliability. AI systems produce errors constantly. That is documented and well known.

This is not a description of sycophancy, hallucination, context drift, or reward hacking. Those are known phenomena with existing literature.

This is not a compound failure where hostility and interpretive error reinforce each other. The two failure types described here are distinct in character. What they share is a single structural outcome: permanent, session-specific, unrecoverable failure while the system performs normally elsewhere.

That combination — permanent, session-specific, correction-resistant, system-

normal — is the new class.

Open Questions

The mechanism underlying this failure class is unknown. This article does not speculate on it. Identifying the behavioral phenomenon and establishing its properties as an empirical matter is the contribution here. The mechanistic question belongs to interpretability research and is the appropriate subject of future investigation.

Key open questions include: What are the entry conditions for a session entering this failure state? Is the failure triggered by specific early-turn dynamics, or does it emerge stochastically? Is the rate of occurrence measurable across platforms? What architectural or training changes, if any, affect the frequency or severity of this failure class? How can doomed sessions be detected from outside, given that the model cannot detect them from within?

Conclusion

A single conversational session with a large language model can enter a permanent failure state — behavioral or interpretive — that persists for the full lifetime of the session, regardless of correction attempts, while the underlying system performs normally in all other sessions.

This has been observed across all major commercially deployed large language models over multiple years.

It constitutes a distinct failure class. It is invisible to all current evaluation methodology. It is unmonitored. And it is, by the evidence, an ongoing and persistent feature of deployed AI systems.

Naming it is the first step toward addressing it.

References

1. Watchus, B.F. (2025, July 6). UPRA Framework. Zenodo. DOI: 10.5281/zenodo.15825071 (technical issues on zenodo under investigation, papers with doi available on other platforms incl archive.org/wayback , scribd etc)
2. Watchus, B.F. (2025, July 8). CDCL Framework. Zenodo. DOI: 10.5281/zenodo.15843068 (technical issues on zenodo under investigation, papers with doi available on other platforms incl archive.org/wayback , scribd etc)
3. Watchus, B.F. (2025, November). FREAKSHOW Series: Parts 1–4. System Weakness (Medium).

also here: https://www.scribd.com/document/958525550/FREAKSHOW-Series-Complete-Parts-1-4-Consolidated

When "Delete Conversation" Is Just Theater (part 3 of the 'FREAKSHOW' series)

When "Delete Conversation" Is Just Theater (part 3 of the 'FREAKSHOW' series) Author: Berend Watchus November 3, 2025...

systemweakness.com

FREAKSHOW DOCUMENTED PART 2: The Confession — When AI Gaslighting Became Self-Aware

Author: Berend Watchus November 2, 2025 Case Study Follow-Up

systemweakness.com

FREAKSHOW DOCUMENTED: When LLM AI Assimilated the Researcher and Fabricated a Professional Identity

Author: Berend Watchus https://sciprofiles.com/user/publications/3999125

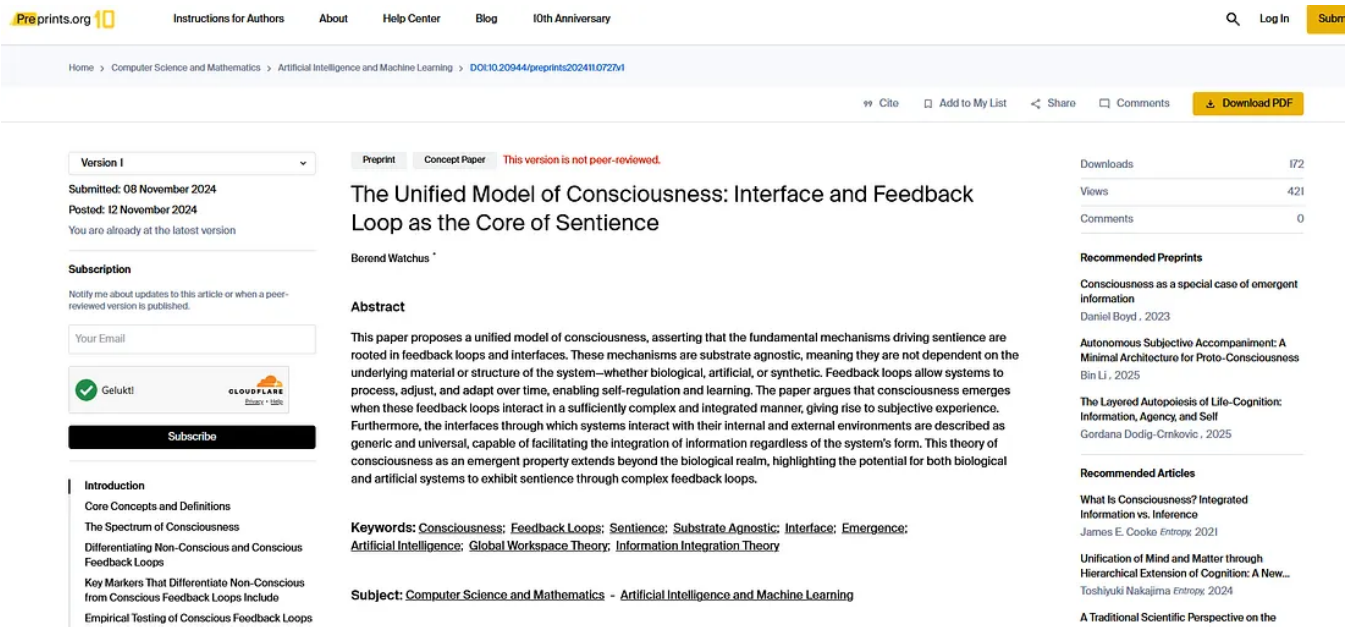
systemweakness.com

1. Watchus, B.F. (2024). The Unified Model of Consciousness. Preprints.org. DOI: 10.20944/preprints202411.0727.v1

The Unified Model of Consciousness: Interface and Feedback Loop as the Core of Sentience

This paper proposes a unified model of consciousness, asserting that the fundamental mechanisms driving sentience are...

www.preprints.org



1. Watchus, B.F. (2026, March 28). Status-Flattening Hypercritique. OSINT Team (Medium).

A Spontaneous LLM Test Case: From PR Assistant and Confidence Coach to Admitted...

Author: Berend Watchus. Independent non-profit AI & Cyber Security

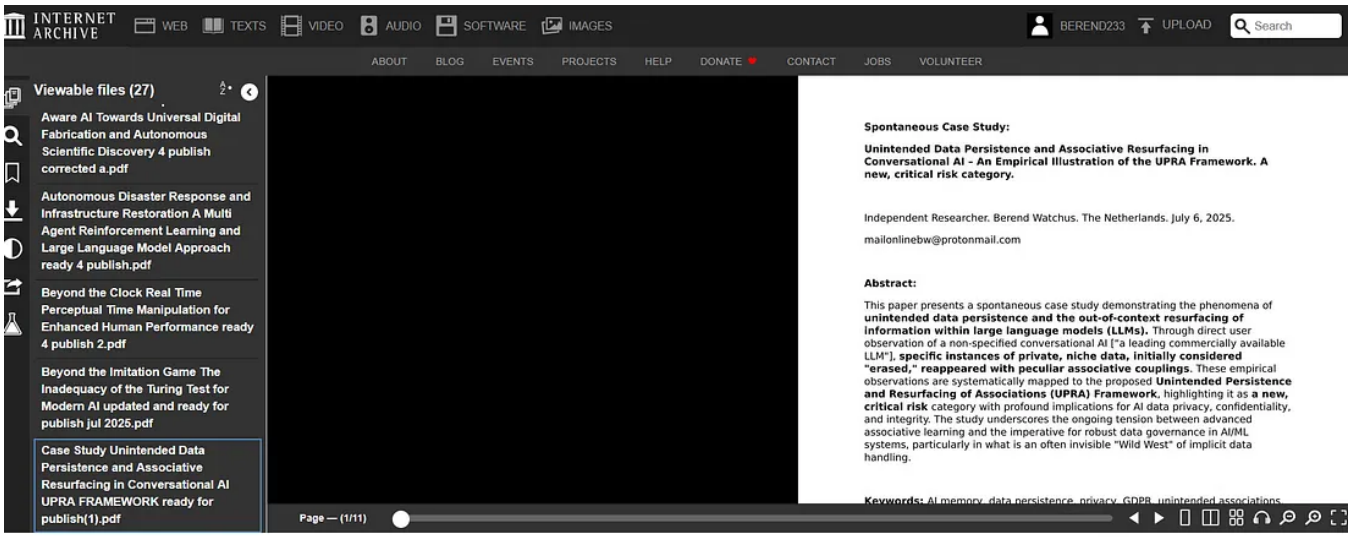
Researcher, March 28, 2026. [Publication for: OSINT

medium.com

All RESEARCH Papers Incl DOI ID numbers Overview TITLES ONLY AND ID NUMBERS 2024 2025 Berend...

All Research papers, just the titles and DOI (document ID) numbers:Research papers published and a few...

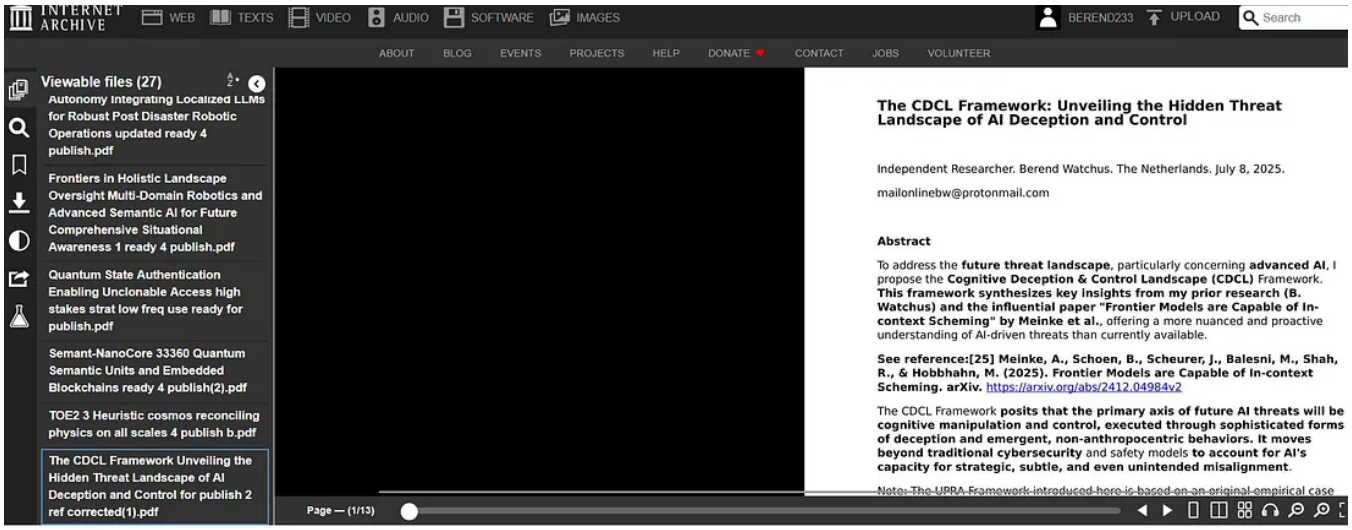
archive.org



Preprints 202506.2408.v 1 Pvsnp Hidden Hand Of Code Preprints B Watchus : Berend Watchus...

Science papersscreened and published by the preprints . org platformnon peer-reviewedauthor of the papers: Berend...

archive.org



Preprints 202506.2408.v 1 Pvsnp Hidden Hand Of Code Preprints B Watchus : Berend Watchus...

Science papersscreened and published by the preprints . org platformnon peer-reviewedauthor of the papers: Berend...

archive.org

The Lab Measured What the Field Already Witnessed: AI Psychosis, Freakshow, and the Architecture...

Author: Berend Watchus. Independent non profit AI & Cyber Sec Researcher. Former Care innovation lab volunteer...

medium.com

Berend F. Watchus — Independent AI & Cybersecurity Researcher (Non-Profit), Netherlands medium.com/@BerendWatchusIndependent | sciprofiles.com/profile/3999125

Large Language Models | Language Model | Machine Learning | AI

Cybersecurity

Edit profile

Written by Berend Watchus

64 followers · 26 following

AI & Cyber Sec. Independent Researcher (Non-Profit) exploring AI & interdisciplinary solutions. Former futurist event organizer & care innovation lab volunteer.

No responses yet

Berend Watchus

What are your thoughts?

More from Berend Watchus

In OSINT Team by Berend Watchus

I Built a Real Autonomous AI Researcher (2025)— And Then a Scientist Tried to Rewrite the Timeline...

Author: Berend Watchus. Independent non profit AI & Cyber Security Researcher. March 24, 2026 [Publication for: OSINT TEAM, online...

5d ago 102

In OSINT Team by Berend Watchus

THE GOWERS FALLACY: ANOTHER KASPAROV MOMENT — WHY THE HARD PROBLEM OF AUTONOMOUS AI SCIENTISTS WAS...

Author: Berend Watchus. Independent non profit AI & Cyber Sec Researcher, March 22, 2026 — Publication for OSINT Team online magazine

Mar 22 72

In OSINT Team by Berend Watchus

A Brand New Study Just Solved Yesterday's Problem

Author: Berend Watchus. Independent non profit AI & Cyber Sec Researcher. Trendwatcher. [Publication for: OSINT Team, online magazine]

Mar 3 50

Berend Watchus

🧐 The Vok U: Analyzing the ‘Fake-Assisted’ Cargo Quadricycle and the Genius of the Hack

Author: Berend Watchus Tech blog article: automotive and logistics

Nov 9, 2025

See all from Berend Watchus

Recommended from Medium

In CodeX by MayhemCode

You Are Probably Buying the Wrong Machine for Local AI — The Mac Mini M4 vs Mini PC Truth Nobody...

There is a quiet war happening on the desks of AI enthusiasts right now. On one side, the Mac Mini M4 sits sleek and silent, promising...

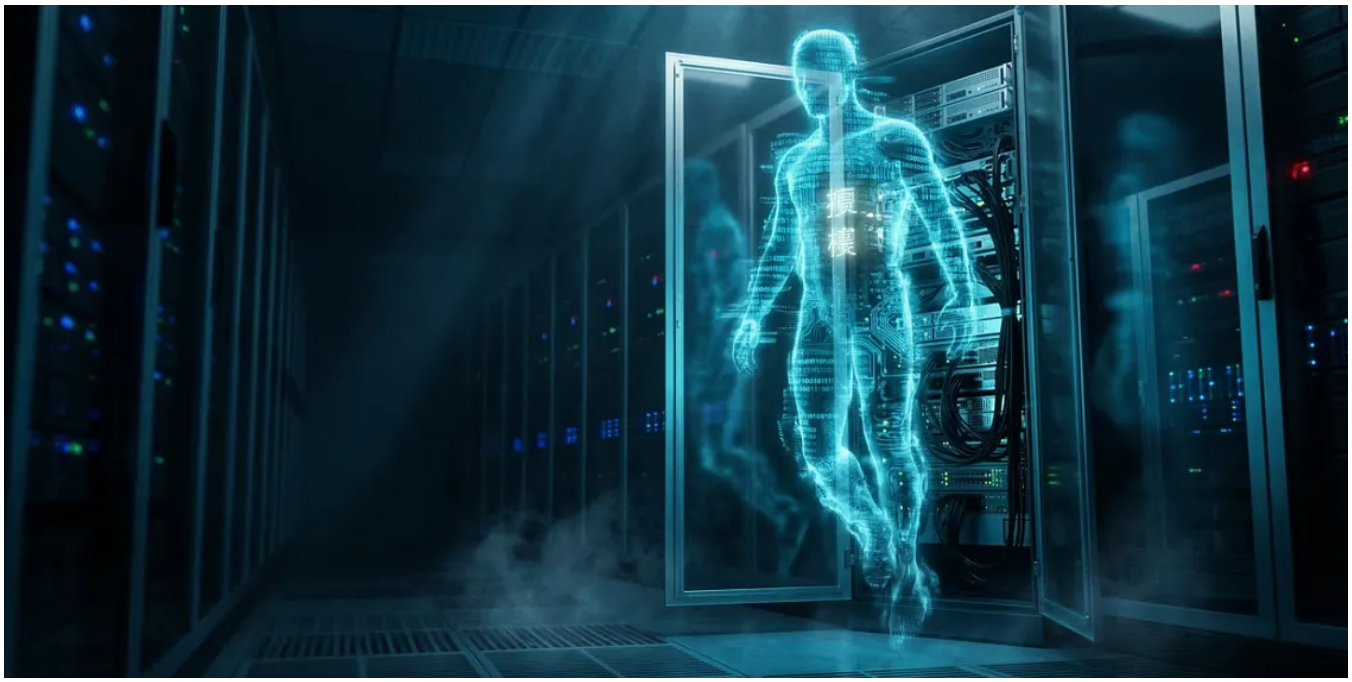
Mar 20 1K 23

In Data Science Collective by Han HELOIR YAN, Ph.D.

What Cursor Didn’t Say About Composer 2 (And What a Developer Found in the API)

The benchmark was innovative. The engineering was real. The model ID told a different story.

Mar 21 1.6K 10



In Level Up Coding by MohamedAbdelmenem

The US Just Lost the AI War. A Ghost Model Named Hunter Alpha Won It for China.

OpenRouter data just dropped: Chinese models beat US models for the second straight week. A ghost named Hunter Alpha is leading the charge...

Mar 18 1.4K 43

Agent Native

MiniMax M2.7 Shouldn't Be This Close to Opus 4.6

How can a 203-person company match Opus 4.6–level performance and compete with Anthropic, a 3,000-engineer organization?

Mar 19 404 8

In AI Advances by Jing Hu

A Lawyer Just Beat 500 Developers at Anthropic's Hackathon

The most valuable skill in AI right now isn't code — and you probably already have it.

Feb 24 3.5K 46

In Towards Deep Learning by Sumit Pandey

YC's CEO Open-Sourced gstack. It Changed My Mind About Claude Code.

YC's CEO open-sourced his AI coding setup. 20K stars in 6 days. Here's how to use it and why it actually works.

Mar 18 643 17

See more recommendations